The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins

David A. C. Beck, Darwin O. V. Alonso, Daigo Inoyama*, and Valerie Daggett†

Department of Bioengineering, Box 355013, University of Washington, Seattle, WA 98195-5013

Edited by Harold A. Scheraga, Cornell University, Ithaca, NY, and approved July 15, 2008 (received for review July 11, 2007)

Here, we compare the distributions of main chain (Φ,Ψ) angles (i.e., Ramachandran maps) of the 20 naturally occurring amino acids in three contexts: (*i*) molecular dynamics (MD) simulations of Gly-Gly-X-Gly-Gly pentapeptides in water at 298 K with exhaustive sampling, where X = the amino acid in question; (*ii*) 188 independent protein simulations in water at 298 K from our Dynameomics Project; and (*iii*) static crystal and NMR structures from the Protein Data Bank. The GGXGG peptide series is often used as a model of the unstructured denatured state of proteins. The sampling in the peptide MD simulations is neither random nor uniform. Instead, individual amino acids show preferences for particular conformations, but the peptide is dynamic, and interconversion between conformers is facile. For a given amino acid, the (Φ,Ψ) distributions in the protein simulations and the Protein Data Bank are very similar and often distinct from those in the peptide simulations. Comparison between the peptide and protein simulations shows that packing constraints, solvation, and the tendency for particular amino acids to be used for specific structural motifs can overwhelm the ''intrinsic propensities'' of amino acids for particular (Φ,Ψ) conformations. We also compare our helical propensities with experimental consensus values using the host–guest method, which appear to be determined largely by context and not necessarily the intrinsic conformational propensities of the guest residues. These simulations represent an improved coil library free from contextual effects to better model intrinsic conformational propensities and provide a detailed view of conformations making up the ''random coil'' state.

coil library | Dynameomics | molecular dynamics | protein folding | host-guest

Protein secondary structure was predicted before the atomic structures of protein were determined (1–3). Conformational preferences of the amino acids were also estimated very early on, beginning with Ramachandran's "map" in 1963, "based solely on repulsive van der Waals" forces in dipeptides (4, 5). Remarkably, these predictions regarding structure and conformational preferences were later largely validated in protein crystal structures (6–8). In the protein folding field, these preferences are seen as both means of excluding regions of conformational space and as driving forces for the formation of secondary structure, both of which limit and bias the necessary search of conformational space required during protein folding.

(Φ,Ψ) dihedral angle distributions are increasingly used to check the validity of structures. Although there can be no doubt about the general tendency of amino acids in globular proteins to populate some regions of (Φ,Ψ) space relative to others, the use of such distributions to judge and refine structures leads to dangerous circular reasoning. That is, (Φ,Ψ) preferences are used as tests of crystal structures, and those very crystal structures are then used to define and support the Ramachandran (Φ,Ψ) angle distributions.

Many experimental studies have addressed amino acid conformational propensities through the host–guest approach in small peptides and proteins whereby the amino acid in question is introduced into a homo- or heteropolymer, and the effect of the perturbation on the "host" is evaluated and attributed to the "guest" amino acid (9). Unlike the behavior of (Φ,Ψ) angles in proteins, short peptides in solution generally do not settle down into one conformational state over time; rather, they represent a dynamic ensemble of conformers in rapid equilibrium (10). As such, care must be taken that the host does not determine the properties of the guest, as discussed in depth below. Determination of intrinsic propensities in small peptides can provide insight into the possible preferred conformations of the polypeptide chain in the unfolded state, which may in turn direct folding (11). In this regard, polyproline II (P_{II}) has received much attention (12).

Computers are now fast enough and molecular dynamics (MD) simulations are reliable enough to address these issues at high resolution to obtain information about the underlying conformational ensembles giving rise to the experimental observables. To this end, we describe exhaustive sampling of amino acid conformations within a solvated end-capped Gly-Gly-X-Gly-Gly pentapeptide to investigate the intrinsic conformational preferences of the 20 naturally occurring amino acids. This and similar peptides have been widely used as models for the unstructured denatured states of proteins (10–15). We begin with control peptide simulations to investigate the effect of neighboring groups and the environment on conformation. We then compare the distributions of (Φ,Ψ) angles in GGXGG peptide simulations with those of experimental structures and a set of MD simulations of 188 native globular proteins with different folds to determine the intrinsic conformational propensities of the amino acids and how these propensities relate to what is observed in proteins.

Results and Discussion

To investigate the intrinsic conformational preferences of the amino acids, MD simulations were performed at 298 K with explicit water for each of the 20 naturally occurring residues in the Gly-Gly-X-Gly-Gly pentapeptide, which is a minimally invasive host. For comparison, control simulations of an Ala dipeptide *in vacuo* and in water are also presented. Ramachandran maps were constructed, and the populations of the conformations were tabulated for all simulations. To draw meaningful conclusions about intrinsic conformational preferences of the amino acids, it is important to obtain ergodic sampling, or in other words, that we reach equilibrium. The simulations were continued long enough to ensure convergence. To put the peptide results in context, we also

Author contributions: D.O.V.A. and V.D. designed research; D.A.C.B., D.O.V.A., and D.I. performed research; D.A.C.B., D.O.V.A., D.I., and V.D. analyzed data; and D.O.V.A. and V.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*Present address: Rutgers University, Department of Medical Chemistry, New Brunswick, NJ 08854.

†To whom correspondence should be addressed. E-mail: daggett@u.washington.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0706527105/DCSupplemental.

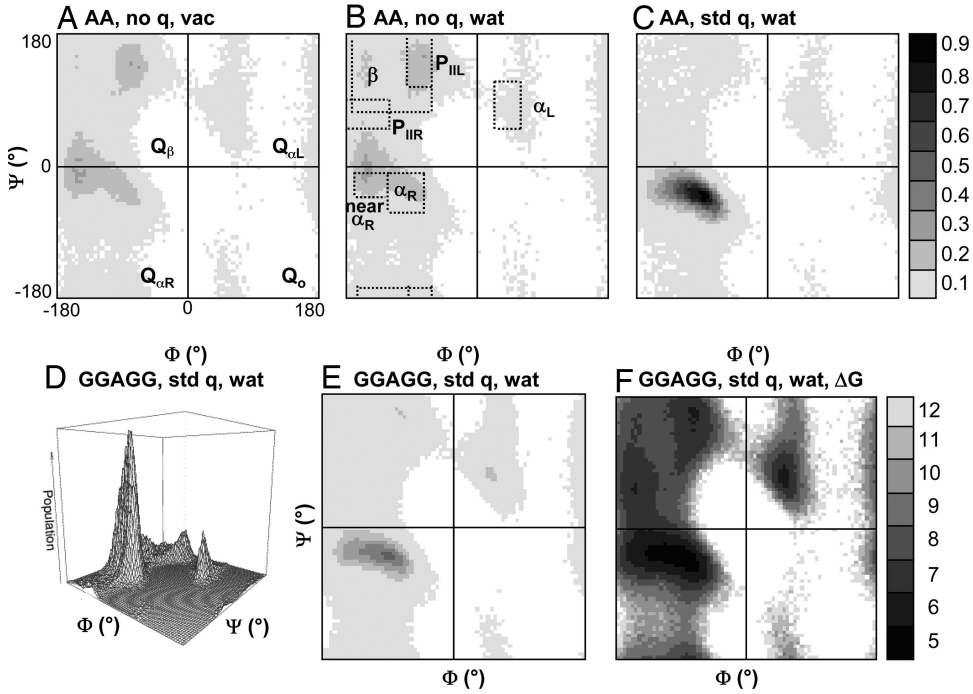

Fig. 1. Ramachandran maps for alanine in different contexts over 100-ns simulations. (*A*) Ala-Ala dipeptide *in vacuo* with all charges set to zero, i.e., only steric effects are considered. (*B*) Ala-Ala dipeptide in water with all peptide and solvent charges set to zero. (*C*) Ala-Ala dipeptide in water with normal charges for peptide and solvent. (*D*) Conformations sampled in a GGAGG simulation. (*E*) Same as *D* but displayed as a contour plot. (*F*) Free energy map for GGAGG calculated from the populations in previous images.

compare conformational distributions of the amino acids in the GGXGG pentapeptide with distributions obtained from simulations of 188 different proteins with varied architectures and experimental distributions from crystal and NMR structures.

Control Simulations. ***AA dipeptide: Effect of environment on conformational properties.*** Because Ramachandran's original work was based on hard sphere space-filling models without solvent, we simulated the Ala-Ala dipeptide without solvent and with all partial charges set to zero. The distributions of (Φ,Ψ) angles in the uncharged solvent-free simulations are similar to the classically accepted Ramachandran preferences (Fig. 1*A*), with the dominant populations occurring in the α- and β-quadrants. The results are similar when water is included, but the system remains uncharged [supporting information (SI) Table S1] and Fig. 1*B*. However, there is a shift toward helical/turn conformers (the α_R and α_L quadrants) when standard, more realistic charges are used and water is included (Fig. 1*C*). The populations are tabulated in Table S1 by quadrant and by the specific conformational region [see labeled regions in Fig. 1 and specific (Φ,Ψ) definitions in *Methods*] for these different dipeptide scenarios. These results show that it is important to consider the environment: The picture provided by uncharged hard sphere *in vacuo* Ramachandran maps may not be accurate for solvated systems, particularly polar solvents (16). It is also noteworthy that there is little change in the distributions upon doubling the sampling from 50 to 100 ns (Table S1).

GGAGG: Sampling behavior. Before analyzing the conformational properties of all 20 aa in depth, sampling and convergence must be addressed. For an end-capped pentapeptide, we assume that each (Φ,Ψ) pair for the three internal residues can exist in four possible conformations with respect to the quadrants in the Ramachandran map, yielding 64 (4^3) possible conformational states. We exclude the first and last residues and the capping groups, because they appear to freely sample conformational space and are not neighbors of the central Ala residue. The fractional population of each of the 64 substates was compared among four independent GGAGG simulations (Fig. S1). Consideration of the conformational properties of three residues instead of just the central residue provides a much more sensitive and robust metric of convergence.

Over short time periods, the conformers sampled in two separate simulations were unlikely to be the same (Fig. S1*A*). However, as sampling approached the ergodic limit, the 64 conformations exhibited equivalent populations in two separate 100-ns trajectories (Fig. S1*B*). One way to measure the sampling convergence is to plot the correlation coefficient (*R*) of the conformer populations between two simulations (Fig. S1*C*). These two trajectories require ≈100 ns for the substate populations to be correlated to >95%, and at 50 ns, they are 90% correlated. Two additional trajectories starting from very different structures converged at the same rate. Based on these findings, all simulations were performed for 100 ns.

GGAGG: Analysis of conformational behavior. Fig. 1 shows a net surface and the accompanying contour plot of the Ramachandran map of Ala for the capped GGAGG peptide (Fig. 1 *D* and *E*). All points sampled by Ala during the simulations are displayed (10^5 points per 100-ns simulation). There are peaks in three different quadrants. The relative peak heights and volumes define Ala's sampling preferences. The same results are obtained in four independent 100-ns simulations. Ala spends ≈60% of its time in the α_R-quadrant, 32% in the β-quadrant, and 8% in the α_L-quadrant (Table 1).

The conformational space sampled by the peptide can be divided more accurately to reflect the underlying local structure adopted (Table 1), as discussed above. Also, given the large number of points and sampling of many different conformations in this small Gly-rich peptide, it can be helpful to convert the population scale into a free energy surface (Fig. 1*F*). The central Ala residue prefers the helical regions of (Φ,Ψ) space, but the bulk of the conformers are shifted from α_R to the elbow region, or near the α_R region, reflective of turns and kinks. Nevertheless, the peptide was very dynamic and covered 51% of the (Φ,Ψ) space during the simulation (using 5° ×

Table 1. Conformational properties (population frequencies) of residue X in GGXGG in 100-ns simulations in water at 298 K

		By quadrant				By specific conformational region						
X	sim #	$Q_{\alpha R}$	Q_{β}	$Q_{\alpha L}$	Q_{o}	α_R	Near α_R	α_L	β	P_{II}	Other	% Coverage
		*	*	*	*	*		*	*	*	*	
Ala	1	58.3	35.8	5.6	0.4	23.5	26.1	4.4	22.3	15.7	19.4	51
Ala	2	59.5	32.9	7.1	0.4	23.6	27.3	5.7	19.0	14.2	19.8	51
Ala	3	61.3	30.4	8.0	0.4	25.1	27.6	6.6	17.8	13.3	18.9	50
Ala	4	54.8	34.7	10.1	0.5	21.6	25.1	8.4	21.2	15.4	19.3	53
Arg	1	65.4	26.6	8.0	0.1	27.9	29.8	7.1	18.8	12.2	14.5	43
Arg	2	69.0	21.2	9.7	0.2	29.5	31.1	8.8	15.2	9.8	14.1	45
Asn	1	80.2	17.9	1.8	0.1	40.3	34.7	1.5	13.6	7.3	8.6	38
Asn	2	81.4	15.2	3.4	0.1	41.2	34.8	3.0	11.1	6.0	8.6	39
Asp	1	77.1	17.8	4.9	0.1	30.9	42.6	4.5	3.9	5.1	14.4	30
Asp	2	74.9	17.9	7.0	0.1	30.9	41.0	6.6	4.7	4.5	14.0	31
Cyh	1	67.2	26.1	6.5	0.2	28.4	33.0	5.8	17.3	11.1	13.0	44
Cyh	2	71.9	24.5	3.4	0.2	31.1	34.6	2.9	16.7	10.1	12.5	43
Gln	1	58.8	26.2	14.8	0.2	26.6	25.5	13.5	18.9	12.0	13.8	44
Gln	2	67.3	23.2	9.3	0.1	29.1	30.4	8.6	15.9	9.8	14.3	43
Glu	1	62.5	30.5	6.2	0.7	31.7	23.6	5.8	22.6	14.4	14.5	45
Glu	2	57.8	32.8	9.3	0.1	29.2	21.6	8.7	24.9	15.1	13.8	44
Gly	1	43.1	9.4	37.3	10.2	16.5	5.5	9.8	7.2	4.8	60.0	76
Gly	2	46.5	9.0	35.5	9.0	18.2	6.7	9.2	6.3	4.3	58.5	76
Gly	3	36.2	9.8	44.5	9.5	13.3	5.1	12.8	7.3	4.8	60.4	75
Gly	4	43.3	10.0	37.8	8.9	16.8	5.5	10.0	7.6	4.9	59.1	76
His	1	65.2	25.4	9.3	0.1	24.7	33.3	8.1	17.7	10.7	14.1	43
His	2	61.0	26.2	12.5	0.2	23.3	31.4	10.8	18.8	10.5	13.8	46
Ile	1	54.5	45.4	0.0	0.0	15.9	29.5	0.0	31.3	21.3	22.0	29
Ile	2	49.9	48.8	1.3	0.0	17.5	23.4	1.2	39.2	27.4	17.6	32
Leu	1	70.1	24.6	5.2	0.1	35.6	27.2	4.3	20.1	12.5	11.9	41
Leu	2	73.7	20.7	5.6	0.1	37.2	29.3	5.0	16.9	10.4	10.8	39
Lys	1	74.2	21.1	4.6	0.1	32.7	33.0	4.2	14.7	9.1	14.1	42
Lys	2	68.1	24.9	6.9	0.1	30.2	30.1	6.2	18.2	11.3	13.7	43
Met	1	66.2	22.6	11.1	0.1	31.1	27.7	10.2	16.4	10.5	13.0	43
Phe	1	71.3	27.1	1.5	0.1	29.1	35.6	1.3	21.1	12.4	11.4	40
Pro	1	20.0	80.0	0.0	0.0	18.3	0.0	0.0	74.2	58.5	7.5	13
Ser	1	54.3	41.4	4.2	0.2	28.4	16.9	3.7	34.0	23.2	15.6	45
Thr	1	50.5	49.3	0.2	0.0	21.5	23.4	0.1	37.5	24.5	15.8	35
Trp	1	72.5	19.9	7.3	0.2	30.8	35.3	6.3	13.4	8.1	12.8	43
Tyr	1	64.5	25.9	9.4	0.3	25.6	32.9	8.2	18.8	10.6	13.0	46
Val	1	56.7	39.3	4.0	0.0	21.9	28.4	3.8	30.4	20.3	14.4	34

See Fig. 1.

5° bins). Detailed breakdowns of the conformations sampled and their populations over the 10^5 structures from each simulation are presented in Table 1. Furthermore, these data are provided for the four independent GGAGG simulations and the results are comparable (Table 1).

The GGAGG peptide has been probed in depth by NMR experiments via proton and heavy atom chemical shifts (13, 14). Chemical shifts were calculated over the simulations by using ShiftX (17) for H_N, H_α, H_β, N_H, C_α, C_β, and C′ for GGAGG are very close to the corresponding experimentally derived values (correlation coefficient $R = 0.999$ over 28 points, 7 atoms × 4 independent simulations). The calculated and experimental values are provided in Table S2.

Another quantity from NMR that provides information regarding the conformations sampled by MD is the vicinal spin–spin coupling constant $^3J_{NHC\alpha}$, which depends on the torsion angle ϕ (18). The Karplus equation is typically used to estimate ϕ from $^3J_{HNC\alpha}$:

$$^3J_{NHC\alpha} = A\cos^2\theta - B\cos\theta + C, \quad [1]$$

where $\theta = |\phi - 60°|$ (19). The coefficients differ depending on the parameterization and folded protein reference set used. The difference between the calculated and experimental values is 0–2.2 Hz (20), and a single coupling constant can map to four different ϕ values. Nevertheless, we have calculated the coupling constant for Ala in the four independent GGAGG simulations by using the Ludvigsen *et al.* (21) parameterization: A = 6.7, B = −1.3, C = 1.5, leading to $\langle ^3J_{HNC\alpha}\rangle = 7.3 \pm 1.8$ Hz. The experimental values for Ala in GGXGG peptides range from 5.8 to 6.1 Hz (13–15). Although our calculated value is high, it agrees with experiment within our uncertainty.

The conformations sampled by Ala in the GGAGG peptide at 298 K by MD are not random and do not fully cover conformational space, even though this small Gly peptide is a model for "random coil" structures. Instead Ala has preferred conformations, namely local α_R, α_L, β, and P_{II} (quantified in Table 1), but these states are dynamic and interconvert readily. These simulations illustrate the potential problems associated with assuming a single or even dominant conformation based on experimental observables over large ensembles, and they call into question fitting procedures to "determine" populations, because different population distributions may reproduce the observables. Given the lack of repeating secondary structure in this peptide, the lack of side chain interactions influencing the local conformations populated, and the convergence reached both within and among independent simulations,

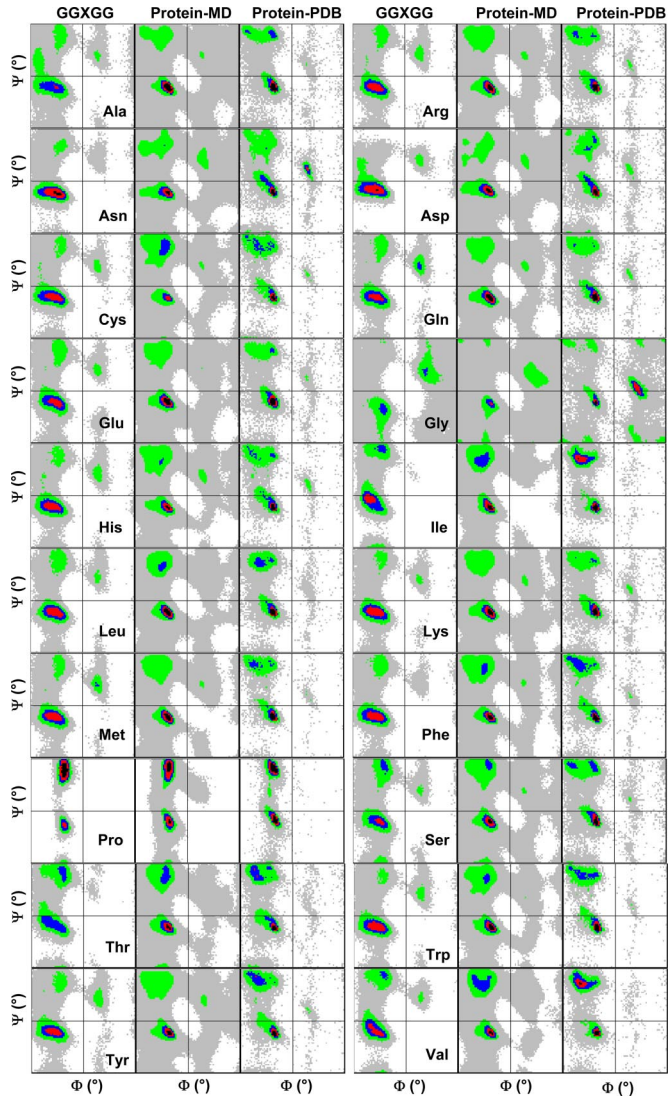

Fig. 2. Ramachandran plots (populations) for the 20 naturally occurring amino acids in different environment contexts: GGXGG peptides, 188 native proteins from our dynameomics set, and 5,626 structures from PDB. All points are plotted over the 100-ns simulations of the peptides and 20-ns simulations of the proteins (the first nanoseconds of all simulations was disregarded). The percentage of points in a bin are colored as follows: 0 ≤ gray < 0.05; 0.05 ≤ green < 0.2; 0.2 ≤ blue < 0.4; 0.4 ≤ red < 0.8: 0.8 ≤ black.

we believe that the distributions obtained reflect the intrinsic conformational propensities of Ala. Consequently, we now expand upon this work to investigate the other 19 amino acids.

Amino Acid Conformations in Different Structural Contexts. ***GGXGG pentapeptides.*** Ramachandran plots were created for each of the amino acids from the distribution of instantaneous Φ and Ψ angles in the GGXGG simulations (Fig. 2). The plots revealed that all of the residues, except for Pro, exhibit a strong propensity for local turns, as reflected in the populations of the α_R and/or near-α_R regions. The α-helix peak is shifted from the strictly defined α-helical region (Fig. 2), which occurs because what we observe is not repeating structure, but the conformational properties of a single residue. The β-sheet, α_L, and P_{II} conformations were also populated to varying degrees (Table 1).

The most highly populated (Φ, Ψ) angles for GGAGG were Φ = (−175 to −50) and Ψ = (−55 to −5), which we define as near-α_R (Figs. 1 and 2). The average Φ and Ψ angles were −110° ± 41° and −12° ± 64°, respectively (calculated by using circular statistics). Consideration of the average in light of the actual conformational space sampled (Table 1 and Fig. 2) illustrates the problem of describing mixed ensembles in this way. As with the GGAGG peptide, we calculated proton and heavy atom chemical shifts over our ensembles for all GGXGG peptides (Table S2). Overall, the agreement between MD and experiment (14) is very good, with a correlation coefficient of 0.9998 for 151 values (Table S3). The agreement for N_H is lower than for the other atoms, possibly because of the difference in pH: MD was done at neutral pH and the experiments were at pH 2.3 in 8 M urea. N_H chemical shifts are known to be very sensitive to pH, temperature, and environment (13). The C_α chemical shifts are very sensitive to conformation, and the good agreement suggests that the MD-generated ensembles are reasonable (R = 0.9879 for the 20 aa, Table S3). Another noteworthy point is that the proton chemical shifts have been determined for the GGXGG peptide series under a variety of pH, temperature, and solvent conditions, and the experimental results are in good agreement with each other and with our MD-derived values ($R \geq 0.996$) (Table S4).

For many of the residues (Asp, His, Phe, Trp, and Tyr), the near-α_R region was the most populated conformational state (Table 1). As mentioned above, Ala sampled 51% of the (Φ,Ψ) space (Table 1). Not surprisingly, the highest value was for Gly at 75% coverage. The lowest was 13% for Pro, followed by ≈30% for Asp and Ile. For comparison, the values for Ala and Gly at 498 K are 68 and 82%, respectively (22).

Free energies were calculated for α_R relative to all other conformations (Table S5). For example, for Ala, the P_{II} content was ≈15% in four independent simulations. Comparison of the populations yields $\Delta G_{PII \rightarrow \alpha R} = -0.3$ kcal/mol. In addition, the free energy for helical conformers can be compared with experimentally derived consensus helical propensities from experimental host–guest studies of different peptide and protein "hosts" (Table S5) (23). Contrary to most experiments, Ala did not have the highest helical propensity; instead, Asn and then Leu displayed the strongest helical propensities in the GGXGG peptides. In fact, 14 of the amino acids had higher helical propensities than Ala.

Comparison of amino acid conformational preferences in peptides and proteins. Fig. 2 shows a comparison of (Φ,Ψ) populations for three systems: (*i*) the GGXGG pentapeptides; (*ii*) MD simulations at 298 K of 188 native proteins with varying architecture from our Dynameomics project (24); and (*iii*) 5,626 structures from the nonredundant Astral40 Protein Data Bank (PDB)-derived database (25). In the native simulations and the PDB, the amino acid in question was not required to have flanking Gly residues. The native-state MD simulations and the experimental structures sample very similar (Φ,Ψ) regions, whereas the pentapeptides can have quite different preferences (Fig. 2). The Pro and Thr distributions are the most similar in the various simulations, or in other words, they are the least sensitive to environmental influence. Most other residues experienced shifts from the pentapeptide distributions, reflecting the importance of context in determining conformation. The large nonpolar residues (for example, Leu, Met, and Trp) showed the greatest difference between the pentapeptide and protein environments (Fig. 2). That these residues are more sensitive makes sense, given that they are subject to conformational constraints due to their burial in hydrophobic cores.

Intrinsic Conformational Preferences of Amino Acids. Most textbook depictions of Ramachandran plots show almost equal populations in the α_R and β quadrants, with weaker populations in the α_L quadrant (Fig. 1*A*). Inclusion of water and realistic partial charges on the atoms shift these distributions away from β-structure in the peptide. The distributions for these same residues in simulations of 188 native proteins and in the PDB, however, have higher β populations, reflecting the need for tertiary contacts to stabilize such conformers (Table 1 and Fig. 2).

Although pentapeptides are not proteins and do not fold to or stay within any one conformational state, they do sample some substates preferentially. The dynamic nature of peptides and the number of possible substates provide for a subtle comparison between simulations and a test of sampling. Different trajectories of the same system are unlikely to sample the same substates at any given time, but averages over longer periods of time must converge. That is, specific substates are favored and sampled preferentially but given a long enough sampling time, the two trajectories should, on average, provide the same populations for a given substate. We found this to be the case provided the simulations were 50–100 ns.

The disparity in (Φ,Ψ) distributions between isolated solvated peptides and in the context of a folded protein is due primarily to direct interactions with water and by packing constraints in proteins. In both cases, the environment has a dramatic effect on the so-called propensities. For example, the free energy for the α_R conformation in the GGXGG simulations is poorly correlated (R = 0.28) with the consensus helical propensity scale of Pace and Scholtz (23). However, the correlation is much better for the MD-generated free energy profiles of the native protein set (R = 0.84, or 0.92 excluding Pro) (Table S5). And, as expected from experimental host–guest studies, Ala has the highest helical propensity in the protein set.

We believe that the GGXGG peptides in water approach the true intrinsic conformational propensities of the amino acids, and the host-guest-like studies in peptides and proteins reflect the ability of a residue to fit within the structural environment provided by the host's scaffold. Likewise, the conformational preferences of peptides are not good predictors of the conformations adopted in proteins. That said, it is interesting that, although the distributions and populations differ when comparing the amino acids, they are not dramatically different. For example, consider the α-helix "breaker" valine. In an isolated pentapeptide, the α-helical state is highly populated by Val (Fig. 2 and Table 1). However, when moving to larger peptides and proteins (Fig. 2), Val is forced out of the helical region because of interactions with neighboring residues. This behavior is context-dependent: Val has no inherent problem adopting helical (Φ,Ψ) values. Similar arguments can be made for other residues.

The "unstructured" or "coil" regions of proteins have been analyzed independently of the more well ordered secondary structure (26–30). These analyses were aimed at obtaining intrinsic conformational preferences of amino acids in contrast to preferences dictated by the role of the amino acids in protein structure, packing and solubility. The exposed portions of the protein lacking regular secondary structure, experience a striking increase in the number of points in the α_L region (30), as observed in our peptide simulations. In agreement with our Val results, Griffiths-Jones *et al.* (30) found that context is critical for β-structure, which is determined primarily by side chain interactions, as found in earlier experiments (31).

Although our pentapeptide distributions are in better agreement with the "coil" regions of proteins, such "coil" libraries still contain conformational biases imposed by the protein. Also, one must be careful how coil libraries are parsed. For example, Jha *et al.* (26) claim that "the backbone preferentially adopts dihedral angles consistent with the polyproline II conformation rather than α or β conformations." Yet, their own data show nearly equal populations of the three conformers for their optimal library in which helix, sheet, turn, and flanking residues are removed: 27.4, 32.9, and 35.5 for α, β, and P_{II} basins, respectively. Furthermore, when only the structured segments are removed (helix and sheet), helical populations are favored: 37.0, 23.2, and 33.2 for α, β, and P_{II}, respectively. Indeed, turns are an important component of denatured states of proteins and warrant inclusion in a coil library. Their removal appears to bias toward more extended segments, thereby increasing the P_{II} population. Overall, it seems most prudent to not ascribe singular dominance to P_{II} based on the coil library; no matter how the structural segments are parsed, Table 1 of Jha *et al.* (26) shows that the combined α_R and β populations are substantially greater than P_{II}, and in fact the three are essentially comparable.

Furthermore, based on agreement between our protein results and host–guest propensities and lack of agreement with the GGXGG peptides and the longstanding use of GGXGG peptides as models for the random coil state in the NMR community, we believe that our simulations provide an improved description of the true intrinsic conformational properties of the amino acids. Based on our results, a shift from the so-called random coil values from GGXGG peptides, which are assumed to reflect the shift from random coil to more ordered structures, actually reflect consolidation of structure from a more complicated ensemble of interconverting nonrandomly populated conformers.

Conclusions

Here, we present the results of MD simulations GGXGG in water at 298 K to investigate the intrinsic conformational properties of the twenty naturally occurring amino acids. Care was taken to ensure that the peptide sampling was exhaustive, resulting in >4 μs of sampling of the peptide (and another 4 μs of native protein dynamics of 188 different proteins for comparison). Our results indicate that the intrinsic conformational preferences long assumed to determine secondary structure are weak. Instead, the effect of neighboring groups, whether consecutive in sequence or brought together in space, plays a critical role in determining the conformational preferences of amino acids in proteins. The intrinsic conformational preferences displayed by these pentapeptides are closer to those observed in less structured regions of proteins, such as those in "coil" libraries. However, even these "coil" distributions are biased by the presence of the protein. Consequently, we have compiled our pentapeptide data and constructed the Structural Library of Intrinsic Residue Propensities, which is available at www.dynameomics.org. Finally, the GGXGG peptides are commonly used as references for the random coil state for interpretation of NMR data. Here, we show that a the difference between the properties of residue X in GGXGG vs. the system of interest is not merely a shift from random coil to more ordered structures but instead reflects consolidation of structure from a complicated ensemble of interconverting nonrandomly populated conformers.

Methods

MD Simulations of Peptides. All peptide, protein, and solvent atoms were explicitly present in all simulations. The peptide/protein and solvent force fields have been presented (32–34). The MD simulations were performed by using *in lucem* molecular mechanics (*il*mm) with an 8-Å force-shifted nonbonded cutoff (35). (Note that simulations using longer cutoffs provide the same results, although convergence takes longer in some cases.) We used the extended (Φ and Ψ = 180°) conformation as a starting structure to avoid bias. At least one simulation for each of the 20 amino acids within GGXGG was performed. Multiple independent simulations were performed for the GGAGG peptide to investigate sampling and convergence. All of the simulations were performed with fixed ionization states to reflect neutral pH (Asp^-, Glu^-, Lys^+, Arg^+, and His^0) with acetylated and amidated N and C termini, respectively. Also, control simulations of a capped Ala dipeptide were performed both *in vacuo* with all partial charges set to zero and in water using standard charges. All simulations were performed for 100 ns at 298 K. All simulations, including the protein simulations below, were performed by using the microcanonical NVE (constant volume, energy, and number of particles) ensemble, which provides Boltzmann sampling of conformers.

MD Simulations of Native Proteins. Ramachandran plots of native protein simulations were obtained from our ongoing Dynameomics project (www.dynameomics.org), in which proteins and domains representing the most common folds (36) are being simulated by using a standard protocol (24) and loaded into a hybrid relational multidimensional database (37, 38). Here, we report data from simulations of 188 different proteins in water at 298 K. The simulations are all at least 21 ns long with a mean simulation time of 30 ns. Details regarding the proteins, the protocols, and validation have been presented (24). These proteins represent ≈70% of the structures in the PDB.

BIOPHYSICS

At the time these data were compiled, 188 proteins of the 1,130 eventual Dynameomics targets were complete and had been thoroughly analyzed. This resulted in a sample size of $>3.76 \times 10^6$ structures, with (Φ,Ψ) angles for 23,535 separate residues at 1-ps resolution. The amino acid composition of this combined pool varied from a high of 2,014 Leu residues [4.0×10^7 (Φ,Ψ) pairs] to a low of 329 Trp residues, or 6.6×10^6 total Trp (Φ,Ψ) pairs. In total, 4.71×10^8 (Φ,Ψ) pairs were calculated and binned for the proteins.

PDB Analysis. Ramachandran plots of experimental protein structures were generated from the Astral40 database. This database contains structures with <40% sequence identity compiled by Chandonia *et al.* (25) (http://astral.berkeley.edu, http://astral.berkeley.edu/pdbstyle-1.65.html); 5,626 structures (files) were used here. In all cases, if multiple structures existed in a file, only the first structure was used. We considered 5,674 total .ent files, and 48 of those would not parse, giving the 5,626 actually used. Those structures generated 989,001 (Φ,Ψ) pairs, and those were separated by amino acid and binned as described below.

Ramachandran Maps. (Φ,Ψ) pairs were put into 72×72, 5° x 5° bins. The plots were scaled by the total number of data points [i.e., (φ, ψ) pairs], so that sampling could be compared between simulations of different length. The fractional population of bin *i* is defined as $P_i = N_i/N_{data}$. The ''coverage'' was quantified by dividing the number of sampled bins by the total number of bins (5,184 bins). Also, by calculating the sum of the population in defined regions of secondary structure and dividing the results by the total population, we determined the frequency of each secondary structure. The defined regions were: α_R: $-100° \leq \Phi \leq -30°$; $-80° \leq \Psi \leq -5°$; near α_R: $-175° \leq \Phi \leq -100°$; $-55° \leq \Psi \leq -5°$; α_L: $5° \leq \Phi \leq 75°$; $25° \leq \Psi \leq 120°$; β: $-180° \leq \Phi \leq -50°$; $80° \leq \Psi \leq -170°$; P_{II}: P_{IIL}, $-110° \leq \Phi \leq -50°$; $120° \leq \Psi \leq 180°$ and P_{IIR}, $-180° \leq \Phi \leq -115°$; $50° \leq \Psi \leq 100°$. Note that the P_{II} region is considered both by itself and as a part of the β-region, so the sum of the populations can be greater than one.

Calculation of Free Energies. The free energy surfaces were generated by calculating the free energy for each bin as: $\Delta G = -RT \ln(P_i/(1 - P_i))$. The free energy of unsampled bins is undefined and not displayed. We generated a helix propensity scale and benchmarked our scale to the experimental values by calculating the free energy of the α_R conformation relative to all others.

ACKNOWLEDGMENTS. This work was supported by National Institutes of Health Grant GM 50789 and by Microsoft Research through Technical Computing @ Microsoft.

1. Pauling L, Corey RB (1951) Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc Natl Acad Sci USA* 37:729–740.
2. Pauling L, Corey RB (1951) Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci USA* 37:235–240.
3. Pauling L, Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci USA* 37:251–256.
4. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99.
5. Ramachandran GN, Sasisekharan V (1968) Conformation of polypeptides and proteins. *Adv Protein Chem* 23:283–438.
6. Lovell SC, *et al*. (2003) Structure validation by Calpha geometry: phi, psi and Cbeta deviation. *Proteins* 50:437–450.
7. Hovmoller S, Zhou T, Ohlson T (2002) Conformations of amino acids in proteins. *Acta Crystallogr D* 58:768–776.
8. Ho BK, Thomas A, Brasseur R (2003) Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci* 12:2508–2522.
9. Von Dreele PH, *et al*. (1971) Helix-coil stability constants for naturally occurring amino acids in water. *Macromolecules* 4:408–417.
10. Bundi A, Wuthrich K (1979) H-1-NMR parameters of the common amino acid residues measured in aqueous-solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH. *Biopolymers* 18:285–297.
11. Firestine AM, Chellgren VM, Rucker SJ, Lester TE, Creamer TP (2008) Conformational properties of a peptide model for unfolded alpha-helices. *Biochemistry* 47:3216–3224.
12. Shi Z, Chen K, Liu Z, Kallenbach NR (2006) Conformation of the backbone in unfolded proteins. *Chem Rev* 106:1877–1897.
13. Plaxco KW, *et al.* (1997) The effects of guanidine hydrochloride on the ''random coil'' conformations and NMR chemical shifts of the peptide series GGXGG. *J Biomol NMR* 10:221–230.
14. Schwarzinger S, Kroon GJA, Foss TR, Wright PE, Dyson HJ (2000) Random coil chemical shifts in acidic 8 M urea: Implementation of random coil shift data in NMRView. *J Biomol NMR* 18:43–48.
15. Merutka G, Dyson HJ, Wright PE (1995) ''Random coil'' 1H chemical shifts obtained as a function of temperature and trifluoroethanol concentration for the peptide series GGXGG. *J Biomol NMR* 5:14–24.
16. Petukhov M, Cregut D, Soares CM, Serrano L (1999) Local water bridges and protein conformational stability. *Protein Sci* 8:1981–1989.
17. Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. *J Biomol NMR* 26:215–240.
18. Karplus M (1959) Contact electron-spin coupling of nuclear magnetic moments. *J Chem Phys* 30:11–15.
19. Wuthrich K (1986) *NMR of Proteins and Nucleic Acids* (Wiley, New York).
20. Pardi A, Billeter M, Wuthrich K (1984) Calibration of the angular-dependence of the amide proton-C-alpha proton coupling-constants, $^3J_{HN,C\alpha}$, in a globular protein - Use of $^3J_{HN,C\alpha}$ for identification of helical secondary structure. *J Mol Biol* 180:741–751.
21. Ludvigsen S, Andersen KV, Poulsen FM (1991) Accurate measurements of coupling constants from two-dimensional nuclear magnetic resonance spectra of proteins and determination of phi-angles. *J Mol Biol* 217:731–736.
22. Scott KA, Alonso DO, Sato S, Fersht AR, Daggett V (2007) Conformational entropy of alanine versus glycine in protein denatured states. *Proc Natl Acad Sci USA* 104:2661–2666.
23. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J* 75:422–427.
24. Beck DA, *et al.* (2008) Dynameomics: Mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. *Protein Eng Des Sel* 21:353–368, 2008.
25. Chandonia JM, *et al.* (2002) ASTRAL compendium enhancements. *Nucleic Acids Res* 30:260–263.
26. Jha AK, *et al.* (2005) Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry* 44:9691–9702.
27. Swindells MB, MacArthur MW, Thornton JM (1995) Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat Struct Biol* 2:596–603.
28. Smith LJ, *et al*. (1996) Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations. *J Mol Biol* 255:494–506.
29. Serrano L (1995) Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. *J Mol Biol* 254:322–333.
30. Griffiths-Jones SR, Sharman GJ, Maynard AJ, Searle MS (1998) Modulation of intrinsic phi,psi propensities of amino acids by neighbouring residues in the coil regions of protein structures: NMR analysis and dissection of a beta-hairpin peptide. *J Mol Biol* 284:1597–1609.
31. Minor DL, Kim PS (1994) Context is a major determinant of beta-sheet propensity. *Nature* 371:264–267.
32. Beck DAC, Daggett V (2004) Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods* 34:112–120.
33. Levitt M, Hirshberg M, Sharon R, Daggett V (1995) Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. *Comput Phys Commun* 91:215–231.
34. Levitt M, Hirshberg M, Sharon R, Laidig KE, Daggett V (1997) Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *J Phys Chem B* 101:5051–5061.
35. Beck DC, Alonso DOV, Daggett V (2000–2008) *In lucem* molecular mechanics (*il*mm) (Computer Program, University of Washington, Seattle).
36. Day R, Beck DAC, Armen RS, Daggett V (2003) A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci* 12:2150–2160.
37. Kehl C, Simms AM, Toofanny RD, Daggett V (2008) Dynameomics: A multi-dimensional analysis-optimized database for dynamic protein data. *Protein Eng Des Sel* 21:379–386.
38. Simms AM, Toofanny RD, Kehl C, Benson NC, Daggett V (2008) Dynameomics: Design of a computational lab workflow and scientific data repository for protein simulations. *Protein Eng Des Sel* 21:369–377.
39. Osapay K, Case DA (1991) A new analysis of proton chemical-shifts in proteins. *J Am Chem Soc* 113:9436–9444.